SNSQ Ontology: A Domain Ontology for SNSs Data Quality

Liwei Zheng
Department of Software Engineering, Computer School
Beijing Information Science and Technology University
No. 35 Beisihuan Middle Road, Chaoyang District, Beijing, 100101, China
e-mail: zlw@bistu.edu.cn

***Abstract*—the advent of online social networks has been one of the most exciting events in this decade. Many popular online social networks such as Twitter, Wechat, Weibo, LinkedIn, and Facebook have become increasingly popular. The consequences of the poor quality of data in a social network are often experienced in everyday life. This paper gives a domain ontology model, SNSQ Ontology, for data quality in the area of social networks. It could be a knowledge base for the quality assessment of the rich and linkage data in the social network. High-quality data would be relevant in the data searching, analyzing and mining. Based on the SNSQ Ontology the strategy for data quality assessment and repair is given. And the co-influence among the four quality dimensions, completeness, consistency, currency, and accuracy, are discussed to guarantee an effective assessment process.**

Keywords-ontology; social network; data quality assessment

I. INTRODUCTION

The advent of online social networks has been one of the most exciting events in this decade. Many popular online social networks such as Twitter, Wechat, Weibo, LinkedIn, and Facebook have become increasingly popular. Many such social networks are extremely rich in content, and they typically contain a tremendous amount of content and linkage data which can be leveraged for analysis [1]. The linkage data is essentially the graph structure of the social network and the communications between entities; whereas the content data contains the text, images, and other multimedia data in the network.

But the consequences of the poor quality of data in a social network are often experienced in everyday life [2]. There are many works in data quality assessment. Pipino provided some principles that can help organizations develop usable data quality metrics in 2002 [3]. 16 data quality dimensions are provided in this work and these dimensions, such as accessibility, completeness, interpretability, are widely used in data quality assessment. Unfortunately, there are few works which have been provided on the quality assessment for the rich and linkage data in a social network [4,5,6,7].

This paper gives a domain ontology model, SNSQ Ontology, for data quality in the area of social networks. Ontology is a formal naming and definition of the types, properties, and interrelationships of the entities that really or fundamentally exist for a particular domain of discourse. It was always created to limit complexity and to organize information in many research fields of computer science. SNSQ Ontology is such an ontology which could be used in SNS data quality assessment as a knowledge base. Besides the concepts and relations, quality constraints are signally provided in SNSQ Ontology, which will be very useful in the quality assessment of SNS data. The strategy for data quality assessment and repair also is given based on SNSQ Ontology. In the assessment process, we found that the quality of data does not equal to the simple sum of qualities from different dimensions. In fact, when the quality of one dimension improved, there might be a quality decreasing in another dimension. So the co-influence among the four quality dimensions, completeness, consistency, currency, and accuracy, are discussed.

Section II gives an introduction about the SNSQ Ontology, including the concepts and relations in SNSs field, the data quality concepts and the quality constraints based on SNSs concepts. Section III provides the strategy for data quality assessment. Section IV discusses the co-influence among the four quality dimensions, completeness, consistency, currency, and accuracy. Section V makes a conclusion.

II. SNSQ ONTOLOGY

Domain ontology represents concepts which belong to a special application domain. Particular meanings of terms applied to that domain are provided by domain ontology. Since domain ontologies represent concepts in very specific and often eclectic ways, they are often incompatible. This presents a challenge to the ontology designer. Different ontologies in the same domain arise due to different languages, different intended usage of the ontologies, and different perceptions of the domain. Therefore, the SNSQ Ontology was designed as concise as possible, and the quality constraints were given in formal. That made it be a model much easier to be understood than some other domain ontologies.

Social Networking systems Ontology (SNO) that models key entities and their relationships typically found in SNSs were proposed to support the access control process of SNS in 2010[8].

Based on SNO, we propose Social Networking systems data Quality Ontology (SNQO) that models entities and their data quality related concepts, relations, and constraints. This specific ontology is used to support the data quality measurement process of SNSs.

The current version of the ontology comprises of 14 concepts and 10 object properties. Figure 1 gives an

overview of SNQO without quality concepts.

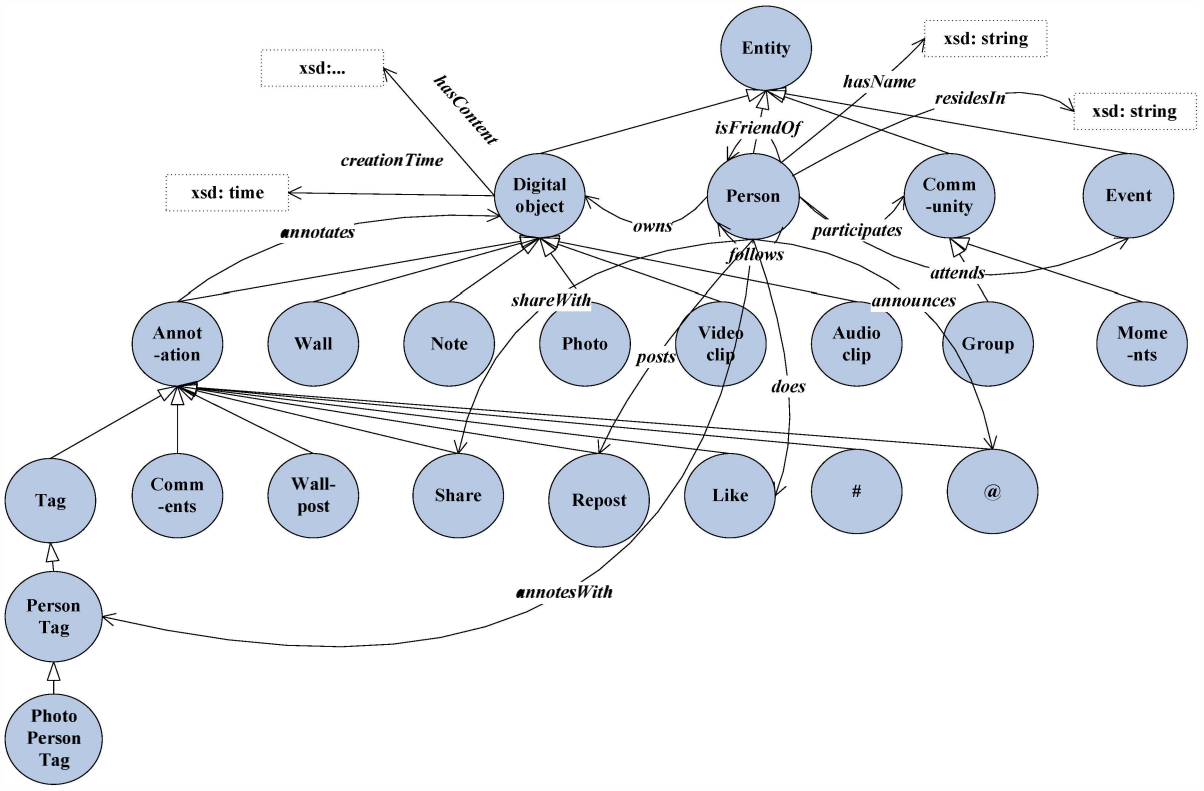

Figure 1. An overview of SNQO without quality concepts.

The Entity concept is the root to all concepts in SNQO, with three immediate descendants: DigitalObject, Person, Community, and Event. The DigitalObject concept models any object with digital, typically visualizable content. The Person concept models human users in the context of SNSs. The Community concept is specialized by subconcepts such as Groups and Moments.

The DigitalObject concept is specialized by subconcepts such as Note, Photo, Video, Audio, Wall, and Annotation.

- The Note concept represents a textual content.
- The Video concept represents a video content.
- The Audio concept represents an audio content.
- The Wall concept models the posting board on the homepage of a person in an SNS, such as the one Facebook provides.
- The Annotation concept represents special digital objects that instead of directly representing content, annotate one object (e.g., a wall, a photo, etc.) using another object (e.g., a textual comment, a person, etc.).

Two objects are related to an annotation object, using properties Annotates and AnnotatesWith, respectively.

Annotation itself is specialized by Comment, Tag, WallPost， follow, unfollow, share, repost, like, friend, unfriend, @(at), and #(hashtag). Comment annotates an object with a note. PhotoPersonTag is a specialized tag that annotates a photo with a person. WallPost annotates a wall with an object, e.g., a photo.

We choose to represent annotation as a concept, rather than a relation, in order to be able to capture more semantics regarding it. For instance, it is usually important to know who has tagged a person in a photo; that might be different from the owner and the tagged person.

Figure 2 shows a piece of instantiated knowledge of SNSQ Ontology.

The knowledge describes Guo's name, where he resides, his friendship with Yu, and events he attends. Guo also owns some Hallowmas photos, which are annotated with Yu by @. Using SNO concepts and relations, more complex semantics can be represented, which is not shown in the sample instantiation. For instance, the @ information mentioned earlier may need to be posted on Guo's wall. For this purpose, a WallPost instance should be created, e.g., wallPost1 has relations Annotates(wallPost1, guoWall) and AnnotatesWith(wallPost1, Yu). Throughout the paper, we use namespace sn to refer to SNO concepts and relations.

There are two types of SNS data, created data and trace data, usually be used in data analysis and mining. For example, the notes, videos, audios uploaded by users are created data; the access time, path, etc. are trace data. The quality of all these data obtained from the Internet should be considered in detail. In this work, there are four quality dimensions are considered, *data consistency, data accuracy, data currency, and data completeness.* Figure 3 gives an overview of SNQO with the four quality concepts.

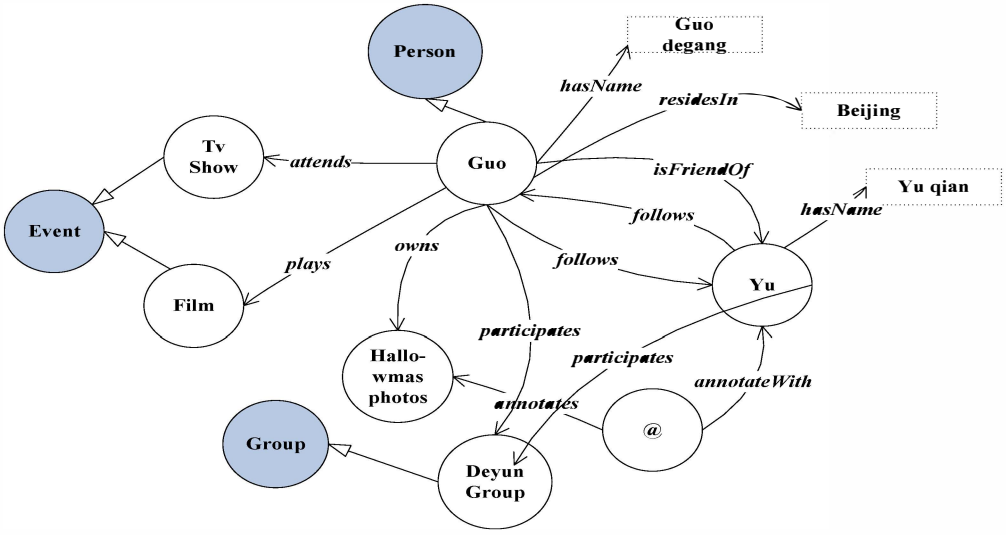

Figure 2. A piece of instantiated knowledge.

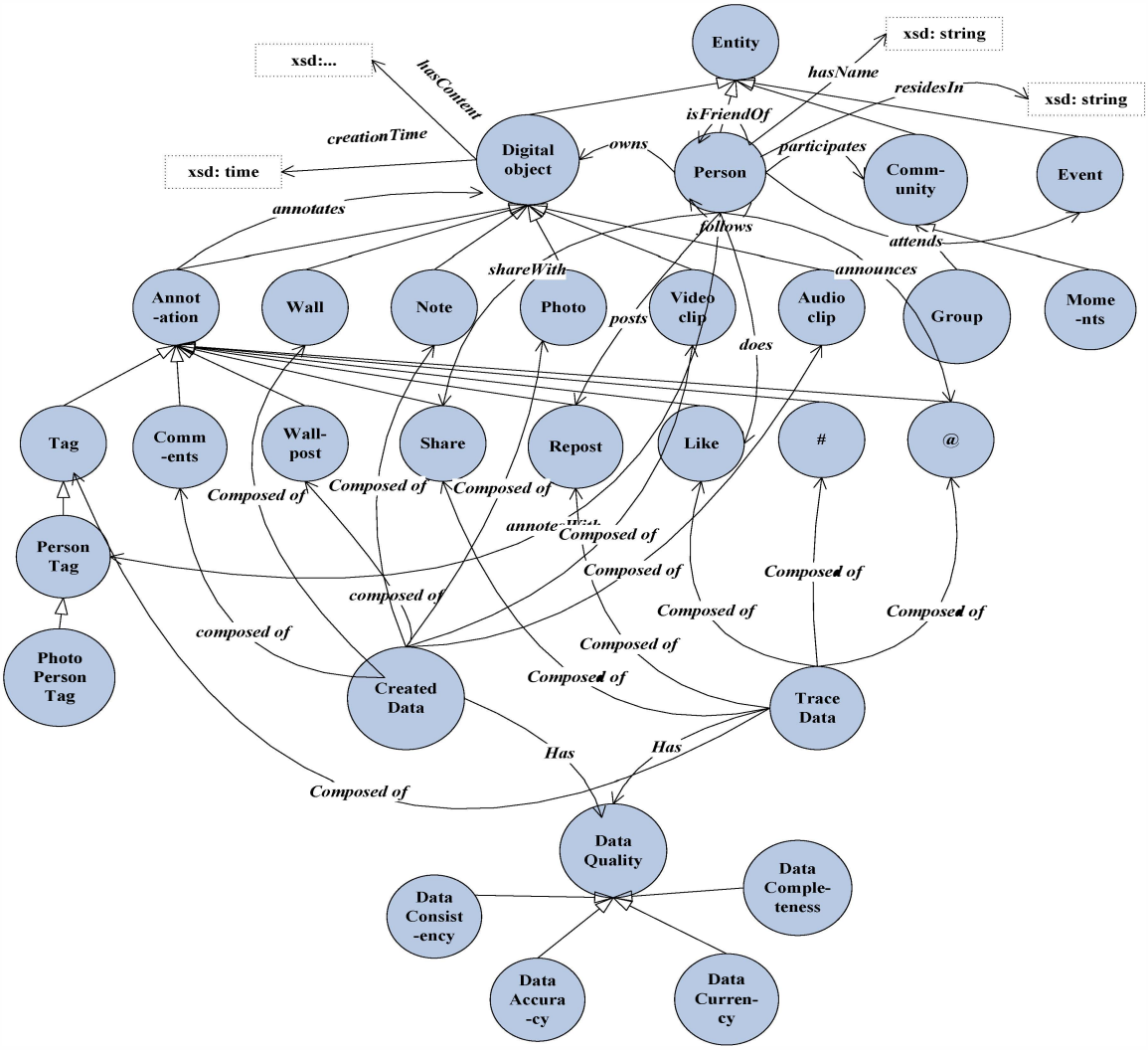

Figure 3. An overview of SNQO with quality concepts.

The four quality dimensions could also have many subconcepts. Figure 4 gives the hasa- hierarchy of the four data quality concepts.

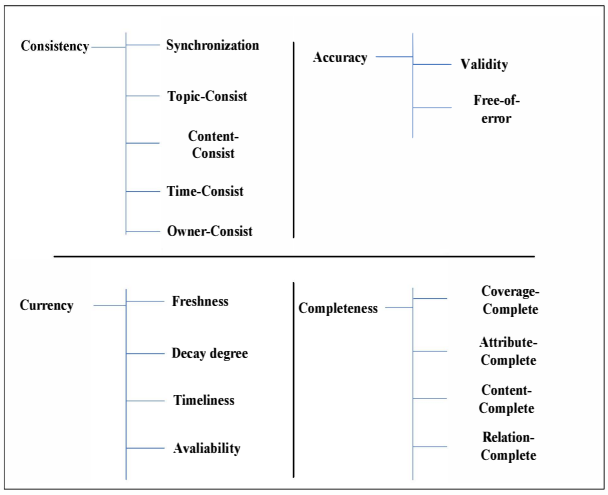

Figure 4. The hasa- hierarchy of data quality concepts.

In classical logic, a theory T is consistency if and only if there is no formula φ such that both φ and its negation $\neg\varphi$ are elements of the set T. That means a consistency theory is one that does not contain a contradiction. Concept "Consistency" represents the consistency constraints on SNS concepts. For the different domain of discourse, "Consistency " could be divided into "Topic-Consist " , "Content-Consist", "Time-Consist" and "Owner-Consist". For example, if a digital object has two owners, that means there exists two persons are the owner of the object, but one resource or object could only have one owner, so an owner inconsistency event occurred.

In the fields of science, engineering, and statistics, the accuracy of a measurement system is the degree of closeness of measurements of a quantity to that quantity's true value. Concept "Accuracy" represents the accuracy of SNS data from two aspects, "Validity" and "Free-of-error". Concept "Validity" mainly concerns on the data validation, while "Free-of -error" mainly concerns about if there is an error.

In the field of data quality, "Currency" is a time-related quality dimension, and it mainly concerns how promptly data are updated. There are also some other "Currency" related sub-concepts. For example, concept "Freshness" concerns the fresh degree of an object, note, or reply. If someone retweets a note which was posted one year ago, then the content retweeted would not be fresh and usable. "Decay degree" represents the data decay degree. When decay degree becomes greater, the freshness becomes less. "Timeliness" expresses how current data are for the task at hand. The timeliness dimension is motivated by the fact that it is possible to have current data that are actually useless because they are late for a specific usage.

"Completeness" can be generically defined as "the extent to which data are of sufficient breadth, depth, and scope of the task at hand" [1]. Four types of "Completeness" are identified. "Coverage-Complete" represents the degree to which SNS concepts and their attributes are not missing. "Attributes-Complete" concerns the completeness of attribute values. "Content-Complete" concerns the completeness of object contents. "Relation-Complete" concerns the completeness of concept relations. The relations of data quality are also complex. Figure 5 shows part of the relations among data quality concepts.

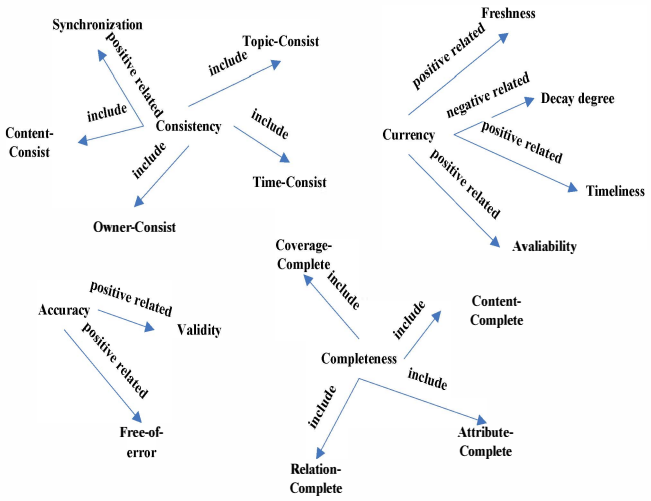

Figure 5. The relations among data quality concepts.

Normally the constraints of quality should be provided before the quality measurement. That means the real quality measurement process needs a more detail and precise definition of quality constraints. So the formally described quality constraints are given as follows.

A. Consistency Constraints

The consistency constraints are given based on the *conditional functional dependencies*(CFDs)[9,10,11]. For example, each blogger has a unique name and unique area at one time. If we have the post time and the blogger's name, we got his area. If there are two area records found, there is a contradiction. Based on these CFDs, we give the definition of SNS concepts consistency constraints.

For $\forall$ data source R in SNS, R could be any digital objects, etc. and $CFDs_R$ is the CFDs defined on R. $\forall c \in CFDs_R$, $c = (R : A \rightarrow B, t)$, for $\forall r \in R$, if any one of the following conditions is true, then r is inconsistency.

$$condition1.r[A] = t[A] \; and \; r[B] \neq t[B]$$

$$condition2.\; if \; r' \in R, r[A] = t[A] \; and \; r'[A] = t[A] \; and \; r[B] = t[B] \; and \; r'[B] \neq t[B]$$

Some examples of the CFDs on the SNS concepts are listed as follows.

$CFD_1 = (R_{\text{person}} : ARea \rightarrow Name, Time, T_{AN})$

$CFD_2 = (R_{\text{note}} : Content \rightarrow Owner, Time, T_{CO})$

$CFD_3 = (R_{\text{Reply}} : Content \rightarrow Replier, Time, T_{\text{CR}})$

$CFD_4 = (R_{\text{Retweet}} : \text{DigitalObject} \rightarrow Retweeter, Time, T_{\text{DR}})$

$CFD_5 = (R_{Group} : Members \rightarrow Group\_name, Group\_owner, Establish-time, T_{\text{MGGE}})$

B. Accuracy Constraints

The Accuracy in SNS concepts mainly concerns three aspects, content accuracy, range accuracy, computation accuracy. Content accuracy focuses on if there is some invalid data or there is some data type error or attributes model error. Range accuracy mainly concerns if the range of record value is enough. Computation accuracy concerns if the field value's precision is enough. Generally, the accuracy problem could be described as follows.

For $\forall$ data source R in SNS, there are n attributes in R and denoted as $A = \{a_1 ..... a_n\}$, suppose the accuracy range set of R is R^*.

$$If \ \ \exists t \in R, t^* \in R^*, a_i \in A,$$
$$and\ value(t[a_i]) \notin value(t^*[a_i])$$
$$then\ t\ is\ not\ accurate.$$

C. Currency Constraints

The currency problem in SNS concepts always concerns if the data record is fresh or new. Whether data is updated in time is the main fact to be considered. For example, if a person finds a discount announcement in someone's note, but the note is posted one month ago. Such a note is meaningless and unuseful. So currency conditions must be given based on the special application background. In a different area, the usable time range is different. These currency conditions could provide the possibility of comparing two records [12].

If *cc* is a given currency condition, Let $\prec_{cc}$ be a compare operator based on cc. For record r1 and r2, r1 $\prec_{cc}$ r2 means r1 is r2 is fresher than r1.

For $\forall r \in R$, if there is never exists r^* which satisfies $r \prec_{cc} r^*$, then r is timeliness or r is fresh enough.

D. Completeness Constraints

The completeness of SNS concepts mainly concerns if there is some null in records if the data size is enough for use and if there are some relations omitted.

For $\forall$ data source R in SNS, there are n attributes in R and denoted as $A = \{a_1 ..... a_n\}$, if any one of the following conditions is satisfied, then there is a completeness error occurred in R.

$condition1.\ \exists r \in R, a_i \in A, value(r[a_i]) = null$

$condition2.\ suppose\ the\ standard\ record\ set\ of\ R\ is\ R^*,$
$A^*\ is\ the\ attribute\ set\ of\ R^*.\ A \cap A^* \neq A^*$

$condition3.\ sizeof(R) < sizeof(R^*)$

The above constraints are a general description of all the data quality constraints on SNS concepts. In the bottom level of SDA-Onto, there are many CDFs and currency conditions drew from real SNS data. And many more different constraints were derived from different areas.

III. Strategy for Data Quality Assessment

Strategies for data quality assessment in the above four dimensions are provided based on the SNSQ ontology in this paper. The data quality assessment process is given as follows.

Process 1. Main data quality assessment

Input: the assessment data set(ADS)

ADS:=<personInfo , createdData , traceData , Tags>

e.g. (Tomas(person),hello world!(note),2013.6.6(time),null)

Output: assessment results in four dimensions

Mostly, the result should be a list of probability values, e.g. each element of ADS would have an Accuracy probability after the assessment process.

ProcessBegin:

Completeness assessment(ADS);

If(Completeness probability<= Completeness threshhold)

Do Completeness repair.

Currency assessment(ADS);

If(Currency probability<= Currency threshhold)

Do Currency repair.

Consistency assessment(ADS);

If(Consistency probability<= Consistency threshhold)

Do Consistency repair.

Accuracy assessment(ADS);

If(Accuracy probability<= Accuracy threshold)

Do Accuracy repair.

ProcessEnd

There are two problems in the above process should be discussed. The first problem is whether the composition of *ADS* could be changed. The answer is yes. Under the different background of data using, the composition of

assessment data set could be different. The quality would be evaluated based on the quality constraints which related to the current data set. The second problem is whether the assessment sequence could be changed. An analysis of the assessment sequence is given at the end of this section. According to this analysis, the given sequence is meaningful and should not be changed. The detailed assessment processes are given as follows.

A. Completeness Assessment

Process 2. Completeness assessment

Input: the assessment data set(ADS)

Output: Completeness assessment results

ProcessBegin:

If(none of the data in ADS has been evaluated)

Do Completeness assessment based on completeness constraints in SNSQ Ontology

Subprocess begin

Suppose there are Ncc completeness constraints in SNSQ Ontology.

For each element in ADS

Suppose there are m constraints are not satisfied.

The completeness probability CPi could be

$$CPi = \frac{Ncc - m}{Ncc} \times 100\%$$

Subprocess end

Record the ADS with completeness assessment values in the instance set of SNSQ Ontology.

If(there exist enough evaluated data which have same structure with current ADS in the instance set of SNSQ Ontology)

Do Training model based on the evaluated data.

//Training method used in this paper is Naïve Bayes.

Apply the model to the unevaluated data, and get the the completeness probability.

ProcessEnd

The above completeness assessment process evaluates ADS data by two different ways. The assessment method based on quality constraints is a classic method and always used for small data scale. In the other side, the model training method using machine learning technique is more useful in the big data environment, e.g. SNS.

B. Currency Assessment

Process 3.Currency assessment

Input: the assessment data set(ADS)

Output: Currency assessment results

ProcessBegin:

If(none of the data in ADS has been evaluated)

Do Currency assessment based on currency constraints in SNSQ Ontology

Subprocess begin

Suppose $\prec_{cc}$ is the time compare operator for ADS , t is the time threshold.

For any element e1 and e2 belongs to ADS, we have

$$e1 \prec_{cc} e2 \Rightarrow t1 > t2$$

For each element in ADS

Suppose ti is the time data of the current element

If(ti>t) the element is fresh and timeliness and the currency probability CuPi is :

$$CuPi = \frac{ti - t}{t - t^0} \times 100\%$$

In which, t^0 is the beginning time of data creating in *ADS.*

Subprocess end

Record the ADS with currency assessment values in the instance set of SNSQ Ontology.

If(there exist enough evaluated data which have same structure with current ADS in the instance set of SNSQ Ontology)

Do Training model based on the evaluated data.

Apply the model to the unevaluated data, and get the currency probability .

ProcessEnd

The currency assessment process is always influenced by the time stamp of data. If the timestamp is missing, the currency assessment would be very difficult. However, we provide a currency assessment method based on time-related constraints which have been used without timestamp. This method would be published soon in our paper titled as a *measurement for social network data currency and trustworthiness*.

C. Consistency Assessment

Process 4. Consistency assessment

Input: the assessment data set(ADS)

Output: Consistency assessment results

ProcessBegin:

If(none of the data in ADS has been evaluated)

Do Consistency assessment based on currency constraints in SNSQ Ontology

Subprocess begin

If(none of the data in ADS has been evaluated)

Do Consistency assessment based on Consistency constraints in SNSQ Ontology

Subprocess begin:

Suppose the constraints set of ADS is CFDs ,let

$N_{cfd} = |CFDs|$

For each element in ADS

Suppose there are m constraints are not satisfied.

The consistency probability CTPi could be

$$CTPi = \frac{N_{nfd} - m}{N_{nfd}} \times 100\%$$

Subprocess end

Record the ADS with Consistency assessment values in the instance set of SNSQ Ontology.

If(there exist enough evaluated data which have same structure with current ADS in the instance set of SNSQ Ontology)

Do Training model based on the evaluated data.

Apply the model to the unevaluated data, and get the Consistency probability.

ProcessEnd

The inconsistency is always imported by wrong input data or computing errors. Mostly, the identification of inconsistency is based on given consistency constraints. It will be very difficult when the consistency constraints are missing. The consistency probability provided based on the prediction of trained model would be an important reference without consistency constraints.

D. Accuracy Assessment

Process 5. Accuracy assessment

Input: the assessment data set(ADS)

Output: Accuracy assessment results

ProcessBegin:

If(none of the data in ADS has been evaluated)

Do Accuracy assessment based on accuracy constraints in SNSQ Ontology

Subprocess begin:

Suppose there are Nac accuracy constraints in SNSQ Ontology.

For each element in ADS

Suppose there are m constraints are not satisifed,that means there are m data items in the current element, which is not accurate.

The accuracy probability Api could be

$$APi = \frac{Nac - m}{Nac} \times 100\%$$

Subprocess end

Record the ADS with accuracy assessment values in the instance set of SNSQ Ontology.

If(there exist enough evaluated data which have same structure with current ADS in the instance set of SNSQ Ontology)

Do Training model based on the evaluated data.

Apply the model to the unevaluated data, and get the accuracy probability .

ProcessEnd

Accuracy always cares about the precision of data. If the accuracy range of data is given, the accuracy assessment process would be very easy. If there is an *ADS* without accuracy range, the machine learning method also could provide a prediction of the accuracy.

IV. Discussion on Co-Influence of Quality Dimensions

The quality of ADS does not equal to the simple sum of qualities from different dimensions. In fact, when the quality of one dimension improved, there might be a quality decreasing in another dimension. There are some relations between data quality dimensions [9].

The repair or improvement of accuracy would make the data more accurate. There is no information missing or destroy. So when the accuracy improved, the completeness will not be changed. The accuracy repair will only change the accuracy range of data; the data value will not be changed. So accuracy repair will not influence the consistency. Similarly, because the data value not changed, the currency will also not be influenced by accuracy repair.

That means the change of accuracy will not influence the other three dimensions.

The repair of completeness will add new values or attributes into the incomplete data. The new parts of data might be inconsistency when they are elicited from different data source. The new coming data might not be fresh enough, so there are also some influences on currency. Similarly, the new data might also be inaccurate or not accurate enough. So the repair of completeness will influence all the other dimensions.

The repair of consistency will not delete data, so it will not influence completeness. The repair of consistency will modify data value, so it will influence the currency and accuracy.

The repair of currency will also not delete data, so it will not influence completeness. But it will modify some value of attributes, so the consistency and accuracy will be influenced.

Figure 6 gives an influence relation graph among the four quality dimensions.

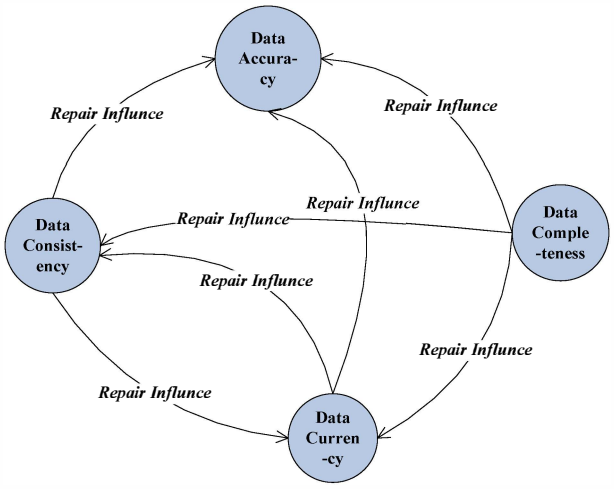

Figure 6. Influence relation graph among quality dimensions.

Based on the influence relations, the assessment sequence could be decided.

- Step1: the completeness assessment and repair
- Step2:the consistency or currency assessment and repair
- Step3:the accuracy assessment and repair

The first assessment is completeness because it will not be influenced by all the other dimensions. The accuracy is the last because it will be influenced by all the others.

The above is a normal discussion on the co-influence among the quality dimensions. In fact, higher quality data set will always get a lower co-influence. And on the contrary, lower quality data set always a higher co-influence. Our assessment sequence might be useful in the lower quality data set assessment and repair.

V. CONCLUSION

This paper gives a domain ontology model, SNSQ Ontology, for data quality in the area of social networks. It could be a knowledge base for the quality assessment of the rich and linkage data in the social network. Based on the SNSQ Ontology the strategy for data quality assessment and repair is given. In the assessment process, we found that the quality of data does not equal to the simple sum of qualities from different dimensions. In fact, when the quality of one dimension improved, there might be a quality decreasing in another dimension. So the co-influence among the four quality dimensions, completeness, consistency, currency, and accuracy, are discussed. The main contributions include:

- Introduces the SNSQ Ontology.
- Provides a strategy for data quality assessment based on SNSQ Ontology.
- Discusses the co-influence among the four quality dimensions, completeness, consistency, currency, and accuracy.

There are many things should be concentrate on in our future work. More quality dimensions would be considered, and an improved model checking method would be given on our ontology.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Funds of China (Grant No. 61402043).

REFERENCES

[1] Carlo Batini, Monica Scannapieca. Methodologies for Data Quality Measurement and Improvement. Data-Centric Systems and Applications. Springer Berlin Heidelberg New York.161-199.2006.

[2] ZHANG Zhi-Gang, JIN Che-Qing, WANG Xiao-Ling, ZHOU Ao-Ying. Discovering Important Locations from Massive and Low-quality Cell Phone Trajectory Data. Journal of software. 2016, 27(7).

[3] Pipino L L, Lee Y W, Wang R Y. Data quality assessment. Communications of the ACM, 45(4): 211-218. 2002.

[4] GUO Zhi-mao, ZHOU Ao-ying. Research on Data Quality and Data Cleaning: a Survey. Journal of software, 13(11), 2077-2082, 2002.

[5] Aebi, D., Perrochon, L. Towards improving data quality. In: Sarda, N.L., ed. Proceedings of the International Conference on Information Systems and Management of Data. Delhi, 273-281.1993.

[6] Wang, R.Y., Kon, H.B., Madnick, S.E. Data quality requirements analysis and modeling. In: Proceedings of the 9th International Conference on Data Engineering. Vienna: IEEE Computer Society,670-677.1993.

[7] Fatemeh Ghorbanpour Alizamini. Data Quality Improvement using Fuzzy Association Rules. In: Proceedings of 2010 International Conference on Electronics and Information Engineering.468-472.2010.

[8] Masoumzadeh A, Joshi J. Osnac. An ontology-based access control model for social networking systems. 2010 IEEE Second International Conference on Social Computing (SocialCom). 751-759.2010.

[9] Ding xiao-ou, Wang Hong-Zhi, Zhang Xiao-Ying, Li Jian-Zhong, Gao Hong. Association relationships study of multi-dimensional data quality. Journal of software. 2016,27(7).

[10] Jin CQ, Liu HP, Zhou AY. Functional dependency and conditional constraint based data repair. Journal of Software, 2016,27(7).

[11] Xu YL, Li ZH, Chen Q, Zhong P. Approach for repairing inconsistency relational data based on possible world model. Journal of Software, 2016,27(7).

[12] Li Mo-Han, Li Jian-Zhong, Gao Hong. Evaluation of Data Currency. Chinese Journal of computers.35(11),2348-2360.2012.